Supplementary Materials

Supplementary Methods

Choosing the appropriate fMRI regressor for the anDDM model (GLMs 1a, b and c)

The attribute-based neural drift diffusion model (anDDM) produces a dynamic accumulation signal that builds over hundreds of milliseconds. This raises a question about the appropriate way to model this signal in the hemodynamic response, which evolves more slowly over 5-10 seconds. To determine the appropriate regressor for GLMs 1a, b, and c, we simulated 5000 instantiations of the anDDM for every subject and trial in Study 2, using a time step of 5 ms. For each subject, we then averaged the 5000 simulations at each time point to produce a single time course of total activity across the two neuronal pools for a given set of trials. We convolved this simulated time course with the canonical form of the hemodynamic response function (HRF) to construct an expected BOLD time series given the inputs. We refer to this as the *ideal BOLD*. We then compared the shape of the ideal BOLD to two different possible instantiations within a traditional GLM analysis in SPM. Version 1 consisted of a parametric modulator of a stick function placed at the onset of the trial, consisting of the sum total activity in the anDDM for each trial, $\sum_{t=1}^{RT} FR_1(t) + FR_2(t)$. Version 2 consisted of a parametric modulator identical to Version 1, but modulating a boxcar function placed at the onset of the trial with duration equal to RT for that trial. Each of these regressors was convolved with the canonical form of the HRF and correlated with the ideal time series to determine the one providing the closest match.

Results suggested that version 2 provided a closer match (Pearson's *r* ranging from .90-.99, average = .96) compared to version 1 (Pearson's *r* ranging from .62-.94, average = .82). Note also that the inclusion of the unmodulated boxcar function with duration equal to the RT on each

trial controls for non-specific activation related to response times that does not build over time in the manner expected based on the anDDM.

Supplementary Results

In the main paper, we focus on the effects of normative vs. hedonistic choice within the dlPFC ROI defined by the conjunction of anDDM-correlated trial-by-trial activity across all three studies. However, in addition to this dlPFC ROI, we identified two other regions, in the dorsal anterior cingulate cortex (dACC, see Figure S3) and left inferior frontal gyrus (IFG)/anterior insula (IFG/aIns, see Figure S4) whose activity correlated with the anDDM across all three studies ($P < .001$, whole brain corrected within each study). Here, we report analogous results on measures of BOLD response in these regions during normative vs. hedonistic choice, for the sake of completeness. These results suggest that our results are a general principle of areas correlating with anDDM response.

dACC response during normative vs. hedonistic choices in Studies 1, 2, and 3

We began by examining whether activity in the dACC correlated with the contrast of normative (generous) vs. hedonistic (selfish) choices in Study 1. As expected, and similar to the dlPFC, this region showed a significantly greater response during generous compared to selfish choices (paired $t_{43} = 3.4825$, $P = .001$, Figure S3d). Similarly, in Study 2, we observed a significant effect of normative goals on the difference in response between normative and hedonistic choices ($F_{2,96} = 13.67$, $P = 5.97 \times 10^{-6}$). Follow-up t-tests confirmed that this was driven by a stronger response in the dACC to normative (generous) choices in Natural trials (paired-$t_{43} = 3.53$, $P = .0009$) as well as significantly stronger response to *hedonistic* choices (paired-$t_{43} = 2.41$,

P = .02) during Partner-focused trials. Finally, we replicated a similar pattern of effects in Study 3, showing a significant influence of normative (i.e., health-focused) goals on the contrast of normative vs. hedonistic choices ($F_{2,96}$ = 3.64, P = .03), which was driven by a stronger response on normative (healthy) choices in the Natural and Taste conditions, and a marginally stronger response on *hedonistic* (i.e., unhealthy) choices during Health Focus trials (paired-t_{43} = 1.96, P = .058).

IFG/aIns response during normative vs. hedonistic choices in Studies 1, 2, and 3

As expected if IFG/aIns response correlates with the anDDM, we observed similar patterns of responding on normative vs. hedonistic choices across all three studies within this region. IFG/aIns showed a significantly greater response during generous compared to selfish choices (paired t_{43} = 3.22, P = .002, Figure S4d). Similarly, in Study 2, we observed a significant effect of normative goals on the difference in response between normative and hedonistic choices ($F_{2,96}$ = 17..66, $P = 2.93 \times 10^{-7}$, Figure S4e). Follow-up t-tests confirmed that this was driven by a stronger response in the dACC to normative (generous) choices in Natural trials (paired-t_{43} = 5.06, $P = 6.57\times 10^{-6}$) as well as significantly stronger response to *hedonistic* (i.e., selfish) choices (paired-t_{32} = 2.66, P = .01) during Partner-focused trials. Finally, we replicated a similar though non-significant pattern of the effects of normative goals in Study 3 ($F_{2,96}$ = .75, P = .39, Figure S4f). However, planned post-hoc comparisons confirmed that activation in the left IFG/aIns was stronger on normative (healthy) choices in the Natural condition (paired-t_{43} = 2.65, P = .01), while activation for this same condition was non-significantly reversed on Health Focus trials (P

= .66). The direct comparison of normative vs. hedonistic choices during Natural vs. Health Focus was also significant (paired-t_{34} = 2.18, P = .04).

Supplementary Figures

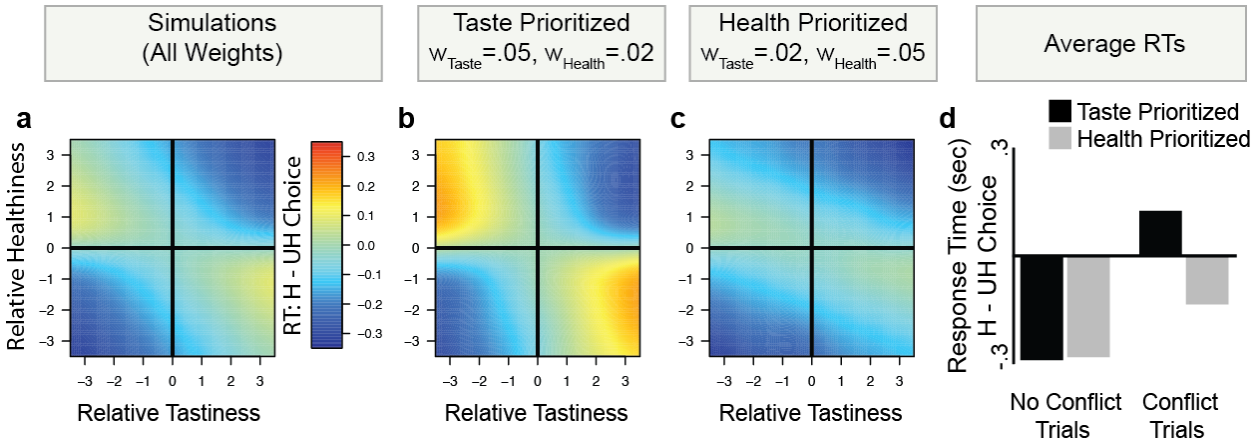

Figure S1. Computational simulations of response time (RT). **(a)** Similar to neural response, model simulations suggest that response times when making normative (i.e., healthy, H) choices instead of hedonistic (i.e. unhealthy, UH) ones (i.e., $RT_H - RT_{UH}$) depends on relative healthiness and tastiness for goal contexts that prioritize both **(b)** hedonism and **(c)** normative goals. Warmer colors indicate longer RTs for healthy choices, indicated by larger differences in $RT_H - RT_{UH}$. **(d)** Average differences in RT for health compared to unhealthy choices (averaging over different options with different attribute values) are displayed for contexts in which health or taste are prioritized, divided as a function of whether relative healthiness and tastiness conflict (i.e., take opposite signs) or do not (no conflict trials). In no conflict trials, on average, healthy choices are easy regardless of whether taste is prioritized (black bars) or health is prioritized (gray bars), indicated by comparatively faster RT_H than RT_{UH}. In conflict trials, however, on average, healthy choices are difficult only in when taste is prioritized (when $w_{Taste} > w_{Health}$), reflected in relatively longer RT_H than RT_{UH}.

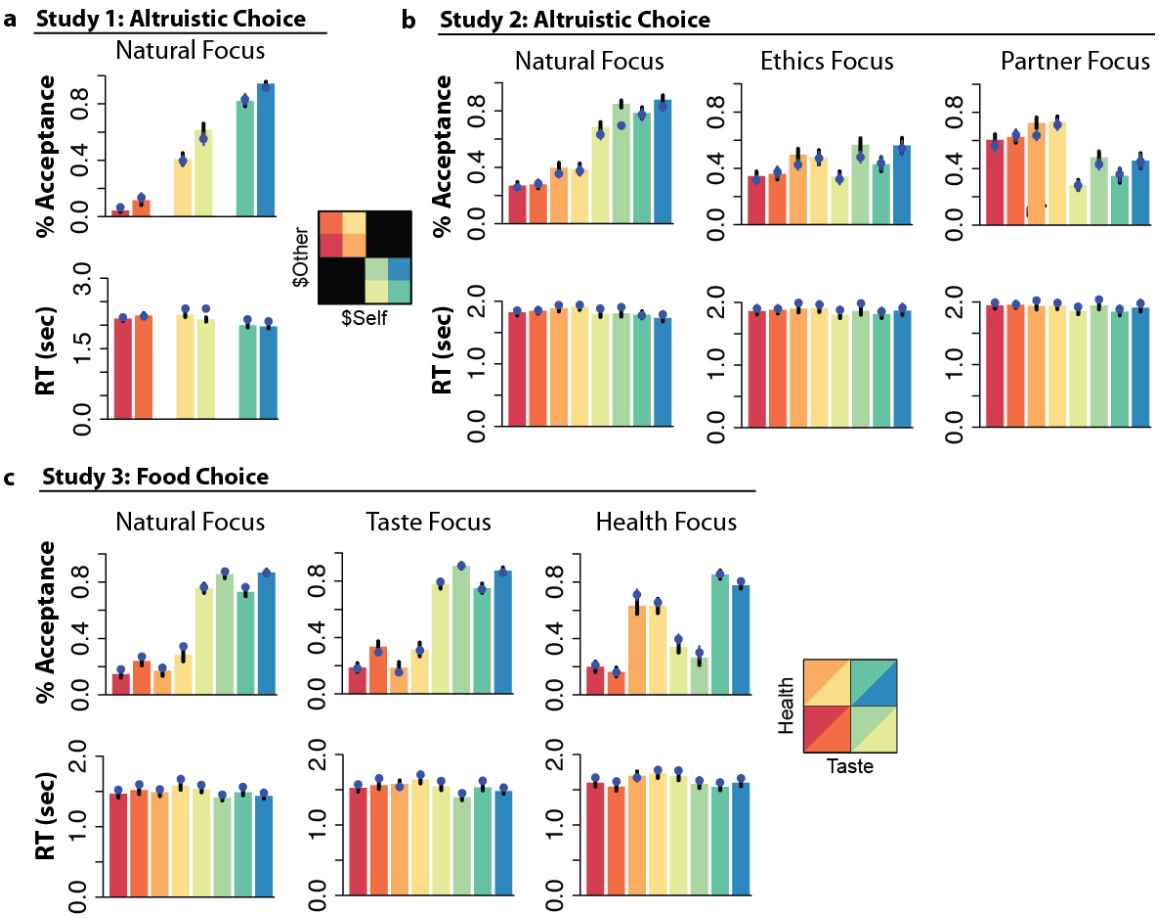

Figure S2. Model fits to behavior. (**a**) Choices and RTs for observed behavior (colored bars) and model simulations (blue dots) for different choice types in Study 1. (**b**) Observed and model-simulated choices and RTs in Study 2, separately by regulatory condition. (**c**) Observed and model-simulated choices and RTs in Study 3, separately by regulatory condition. Error bars show standard error of the mean.

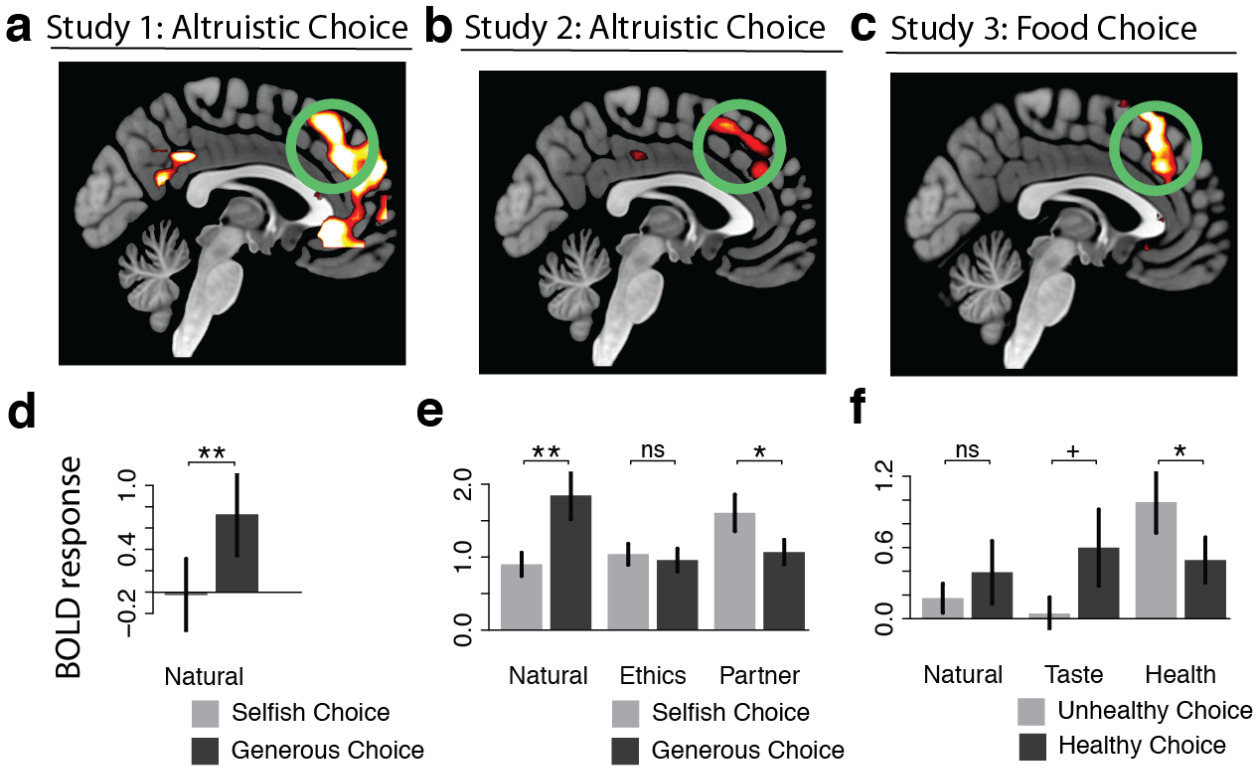

Figure S3. BOLD responses in the anterior cingulate cortex during self-control dilemmas. Top: Trial-by-trial BOLD response in the dACC correlates with predicted activity of the anDDM across three separate studies, including during both altruistic choice **(a, b)** and during dietary choice **(c)**. All maps thresholded at $P < .001$ uncorrected for display purposes. Bottom: Within the dACC ROI defined by the three-way conjunction of anDDM response across all studies, BOLD response during normative choice (black) vs. hedonistic choice (light gray) when attributes conflict, in **d**) Study 1 for all trials, as well as in **e**) Study 2 and **f**) Study 3 as a function of regulatory goals. As predicted, normative choices activate the dACC, but only when goals result in a greater weight on hedonistic than normative attributes. + $P < .05$, one-tailed; * $P < .05$; ** $P < .01$.

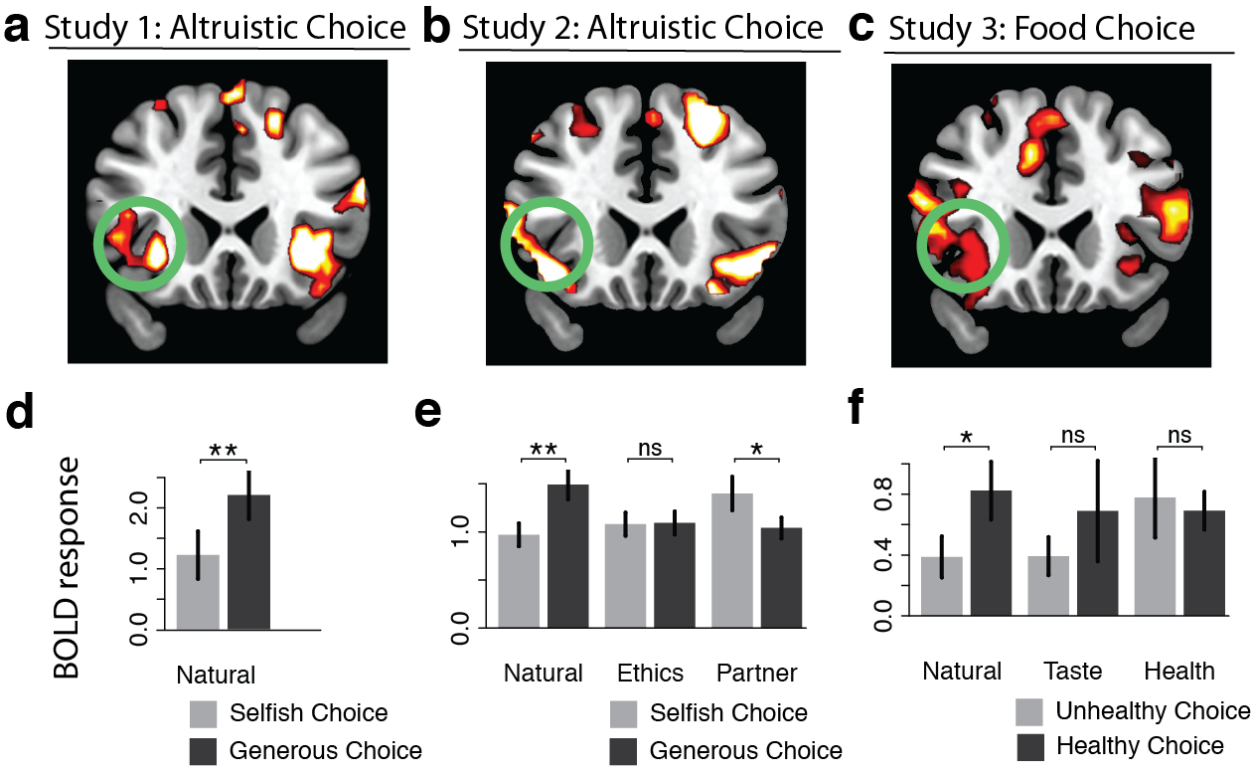

Figure S4. BOLD responses in the inferior frontal gyrus (IFG)/anterior insular cortex during self-control dilemmas. Top: Trial-by-trial BOLD response in the IFG/insula correlates with predicted activity of the anDDM across three separate studies, including during both altruistic choice **(a, b)** and during dietary choice **(c)**. All maps thresholded at $P < .001$ uncorrected for display purposes. Bottom: Within the IFG/insula ROI defined by the three-way conjunction of anDDM response across all studies, BOLD response during normative choice (black) vs. hedonistic choice (light gray) when attributes conflict, in **d**) Study 1 for all trials, as well as in **e**) Study 2 and **f**) Study 3 as a function of regulatory goals. As predicted, normative choices activate the IFG/insulas, but only when goals result in a greater weight on hedonistic than normative attributes. + $P < .05$, one-tailed; * $P < .05$; ** $P < .01$.

Table S1. Estimated Model Parameters

Parameter	*A priori* constraints	Study 1	Study 2, Natural	Study 2, Ethics	Study 2, Partner	Study 3, Natural	Study 3, Taste	Study 3, Health
w_{Self}	-.5 to +.5	.0036±.0011	.0073±.0035[a]	.0061±.0047[a]	.0037±.0065[b]	-	-	-
w_{Other}	-.5 to +.5	.0008±.0015	.001±.0038[a]	.0041±.0045[b]	.0051±.0038[b]	-	-	-
$w_{Fairness}$	-.5 to +.5	.0008±.001	.0017±.0033[a]	.0053±.0046[b]	.0024±.0035[a]	-	-	-
w_{Taste}	-.5 to +.5	-	-	-	-	.0074±.0027[a]	.0077±.0029[a]	.002±.0028[b]
w_{Health}	-.5 to +.5	-	-	-	-	-.0002±.0018[a]	-.0008±.0018[a]	.0055±.0034[b]
B	0 to +1.0	.3181±.1425	.2773±.1373[a]	.3628±.1453[b]	.4062±.1586[b]	.1691±.0501[a]	.1821±.0616[a,b]	.2009±.0819[b]
ndt	0 to +2.0s	.8002±.215	.5989±.2219[a]	.4835±.1448[b]	.4859±.1322[b]	.5442±.1321[a]	.5397±.1361[a]	.5399±.1589[a]
ζ	0 to +2.0	.583±.3034	.5531±.3086[a]	.7469±.2814[b]	.7592±.2806[b]	.4102±.0768[a]	.4093±.0865[a]	.3958±.0848[a]
γ	+1.0 to +3.0	1.8979±.3575	2.0148±.3881[a]	2.2043±.3744[a,b]	2.2952±.3467[b]	1.6435±.1279[a]	1.654±.1578[a]	1.6802±.1455[a]

Note. Parameter values were estimated using a Differential-Evolution Markov Chain Monte Carlo method developed by Holmes and Trueblood[1]. Parameters beginning with *w* indicate weighting parameters applied to different attributes (Studies 1 and 2: proposed payoff to self vs. the default, proposed payoff to other vs. the default, and fairness [|$Self - $Other|]; Study 3: tastiness and healthiness vs. the default). *B*: choice-defining threshold. *ndt*: non-decision time. ζ: lateral inhibition parameter from one neuronal pool onto the other. γ: auto-excitation parameter from a neuronal pool onto itself. *A priori* constraints on the parameters, determined based on previous work and on theoretical limits, restricted them to the range indicated. In Studies 2 and 3, columns indicated by different subscripts differ significantly from each other at $P < .05$, corrected for multiple comparisons.

Table S2. Neural correlates of the attribute-based neural drift diffusion model across studies

Region	BA	Cluster Size	Z score	x	y	z
Study 1 (GLM 1a)						
L Dorsal Anterior Cingulate	6/8/32	235	4.89	-6	27	42
L Inferior Frontal Gyrus	47	271	5.01	-33	27	-6
R Inferior Frontal Gyrus	47	175	4.87	39	27	-6
L Dorsolateral Prefrontal Cortex	45/46	60	4.32	-57	21	24
L Supplementary Motor Area	6/8	142	4.23	-21	12	57
R Inferior Parietal Lobule	40	319	5.76	54	-66	36
L Inferior Parietal Lobule	40	281	5.51	-48	-78	33
Study 2 (GLM 1b)						
R Dorsal Anterior Cingulate	6/8/9/32	936	5.27	-3	35	46
L Inferior Frontal Gyrus	45/47	373	4.82	-45	32	-8
R Inferior Frontal Gyrus	47	268	5.06	39	23	-11
L Dorsolateral Prefrontal Cortex	45	7†	3.6	-57	20	22
L Middle Frontal Gyrus	6/8	293	4.52	-24	20	52
L Posterior Cingulate Cortex	31	100	5.12	-6	-40	34
L Middle Temporal Gyrus	21	38	4.07	-60	-31	-8
L Inferior Parietal Cortex	39	285	5.57	-39	-70	40
R Occipital Cortex		120	4.9	42	-73	34
Study 3 (GLM 1c)						
R Dorsal Anterior Cingulate	6/8/9/32	472	5.28	9	23	40
R Dorsolateral Prefrontal Cortex		684	5.21	54	23	19
Inferior Frontal Gyrus		*	4.11	33	17	-11
L Dorsolateral Prefrontal Cortex		671	5.23	-51	20	19
Inferior Frontal Gyrus	47	*	4.43	-33	26	-5

Note. Regions are reported at a voxel-level of $P < .001$, uncorrected and a whole-brain cluster corrected level of $P < .05$, unless otherwise noted. * Distinct peak within larger cluster. † Significant at $P < .05$, small-volume corrected within a 10-mm spherical region of interest centered on the left dlPFC.

Table S3. Differences in neural response for virtuous vs. hedonistic choices

Region	BA	Cluster Size	Z score	x	y	z
Study 1, Generous vs. Selfish (GLM2a)						
anDDM Regions						
L Dorsomedial Prefrontal Cortex	9/32	86	4.03	-3	33	36
R Dorsolateral Prefrontal Cortex	44/45	24	3.87	54	12	21
L Dorsolateral Prefrontal Cortex	45/46	18*	3.06	-45	12	18
R Inferior Frontal Gyrus	47	23	4.12	30	21	-12
L Inferior Frontal Gyrus	47	13	3.73	-42	39	-3
L Inferior Parietal Lobule	40	14	3.56	-60	-54	39
Other Regions						
No regions significant						
Study 2, Generous vs. Selfish, Natural Focus trials only (GLM2b)						
anDDM Regions						
L Dorsomedial Prefrontal Cortex	9/32/24	2225	5.21	-3	11	67
Dorsomedial Prefrontal Cortex		**	4.79	-9	38	37
Dorsolateral Prefrontal Cortex		**	3.85	-42	14	31
R Dorsolateral Prefrontal Cortex	46	38	3.68	57	23	25
L Inferior Frontal Gyrus	47	381	5.21	-42	20	-8
R Inferior Frontal Gyrus	47	10	3.47	33	17	-11
R Inferior Parietal Lobule	40	20	3.66	48	-37	46
L Inferior Parietal Lobule	40	180	4.42	-39	-67	46
L Inferior Parietal Lobule	40	18	3.59	-57	-37	46
Other Regions						
L Mid-Cingulate Cortex	24	30	4.33	-3	-4	31
R Posterior Cingulate Cortex	31	54	3.79	12	-40	31
R Inferior Parietal Lobule	40	32	3.76	48	-58	46
L Lingual Gyrus	18	37	3.64	-9	-73	1
R Cerebellum		21	3.57	0	-52	-23
L Frontal Pole	10	11	3.38	-9	62	13
Study 2, Generous vs. Selfish, Ethics Focus trials only (GLM2b)						
No regions significant						

Study 2, Generous vs. Selfish, Partner Focus trials only (GLM2b)						
anDDM Regions						
L Dorsomedial Prefrontal Cortex	24	54	-3.64	-3	41	22
R Dorsolateral Prefrontal Cortex	46	16*	-3.81	57	29	22
R Inferior Frontal Gyrus	47	47*	-3.36	36	23	-2
Other Regions						
No regions significant						
Study 3, Healthy vs. Unhealthy, Natural Focus conflict trials only (GLM2c)						
anDDM Regions						
L Dorsomedial Prefrontal Cortex	9	23*	3.82	-12	29	37
L Dorsolateral Prefrontal Cortex	46	7*	3.02	-48	26	16
L Inferior Frontal Gyrus	47	21*	3.09	-27	20	-11
R Inferior Frontal Gyrus	47	14	3.99	30	20	-8
Other Regions						
R Frontal Pole		38	4.19	9	62	4
R Orbitofrontal Cortex		16	3.6	39	41	-5
Study 3, Healthy vs. Unhealthy, Taste Focus conflict trials only (GLM2c)						
anDDM Regions						
R Dorsomedial Prefrontal Cortex	9	8*	3.02	6	23	46
Other Regions						
No regions significant						
Study 3, Healthy vs. Unhealthy, Health Focus conflict trials only (GLM2c)						
No regions significant						

Note. Regions are reported at a voxel-level threshold of $P < .001$, uncorrected, and a minimum volume of $k = 10$ voxels, unless otherwise noted. * Significant at $P < .005$, uncorrected, reported for completeness. anDDM regions are defined by their correspondence with predictions of the attribute-based neural drift diffusion model (anDDM, see Table S2).

References

1. Holmes, W.R. & Trueblood, J.S. Bayesian analysis of the piecewise diffusion decision model. *Behav. Res. Methods* **50**, 730-743 (2018).